CSE 332

INTRODUCTION TO VISUALIZATION

THE VIEWS OF EDWARD TUFTE
AND SOME OTHERS

KLAUS MUELLER

COMPUTER SCIENCE DEPARTMENT
STONY BROOK UNIVERSITY
<table>
<thead>
<tr>
<th>Lecture</th>
<th>Topic</th>
<th>Projects</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Intro, schedule, and logistics</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Basic tasks and data types</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Data sources and preparation</td>
<td>Project 1 out</td>
</tr>
<tr>
<td>4</td>
<td>Notion of similarity and distance</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Data and dimension reduction</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Visual bias</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Introduction to D3</td>
<td>Project 2 out</td>
</tr>
<tr>
<td>8</td>
<td>Visual perception and cognition</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Visual design and aesthetic</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Cluster analysis</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>High-dimensional data – projective methods</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>High-dimensional data – scatterplot displays</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>High-dimensional data – optimizing methods</td>
<td>Project 3 out</td>
</tr>
<tr>
<td>14</td>
<td>Visualization of spatial data: volume visualization intro</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>Visualization of spatial data: raycasting, transfer functions</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>Illumination and isosurface rendering</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>Midterm</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>Scientific visualization</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>How to design effective infographics</td>
<td>Project 4 out</td>
</tr>
<tr>
<td>20</td>
<td>Principles of interaction</td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>Midterm discussion</td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>Visual analytics and the visual sense making process</td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>Visualization of graphs and hierarchies</td>
<td>Project 5 out</td>
</tr>
<tr>
<td>24</td>
<td>Visualization of time-varying and streaming data</td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>Maps</td>
<td></td>
</tr>
<tr>
<td>26</td>
<td><strong>Tufte, memorable visualizations, visual embellishments</strong></td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>Evaluation and user studies</td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>Narrative visualization, storytelling, data journalism, XAI</td>
<td></td>
</tr>
</tbody>
</table>
Seminal Books by Edward Tufte

Standard literature for every visualization enthusiast

Edward Tufte

Well recognized for his writings on information design
• a pioneer in the field of data visualization
• taught information design at Princeton University
• now a professor at Yale University

Popularized concept of “small multiples”
• aka trellis chart or panel chart
• similar charts of same scale + axes
• allows them to be easily compared
• use multiple views to show different partitions of a dataset
E. Muybridge’s Horses in Motion (1886)

- proofed for the first time that horses CAN have all 4 legs in the air
- work was also foundational to the development of the motion picture
FA Walker’s census charts (1870)

- Population is broken down by state and then occupation, including a count of those attending school
- Also has tree maps!
Edward Tufte

Also popularized “sparklines”

- small integrative visualizations

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Bot</th>
<th>Ask</th>
<th>Last</th>
<th>Change</th>
<th>Chart</th>
<th>Volume</th>
<th>High</th>
<th>Low</th>
<th>Value Change</th>
<th>Value</th>
<th>Gain</th>
</tr>
</thead>
<tbody>
<tr>
<td>DCJ</td>
<td>93/4</td>
<td>69/13</td>
<td>69/3/4</td>
<td>5/4</td>
<td>1+1/4</td>
<td>10,210</td>
<td>90/10</td>
<td>69/1</td>
<td>+1.41%</td>
<td>259</td>
<td>17,950</td>
</tr>
<tr>
<td>CPG</td>
<td>48/7/6</td>
<td>48/9/6</td>
<td>48/13/6</td>
<td>-13/8</td>
<td>-13/8</td>
<td>25,620</td>
<td>51/14</td>
<td>46/1</td>
<td>-1.85%</td>
<td>-81</td>
<td>4,804</td>
</tr>
<tr>
<td>SOTI</td>
<td>26/1/4</td>
<td>26/3/8</td>
<td>25/3/8</td>
<td>+1/2</td>
<td>+1/2</td>
<td>504/60</td>
<td>27/36</td>
<td>25/58</td>
<td>+1.93%</td>
<td>250</td>
<td>13,186</td>
</tr>
<tr>
<td>COMS</td>
<td>46/1/2</td>
<td>46/9/16</td>
<td>45/9/16</td>
<td>-25/2</td>
<td>-25/2</td>
<td>3,131</td>
<td>47/15/6</td>
<td>45/3</td>
<td>-1.85%</td>
<td>-102</td>
<td>6,053</td>
</tr>
<tr>
<td>LU</td>
<td>111/5/6</td>
<td>111/11/1</td>
<td>111/9/6</td>
<td>+1/9/6</td>
<td>+1/9/6</td>
<td>5,124</td>
<td>112/5/6</td>
<td>110</td>
<td>-1.42%</td>
<td>73</td>
<td>5,570</td>
</tr>
<tr>
<td>YHOO</td>
<td>388/1/16</td>
<td>388/1/6</td>
<td>388/1/2</td>
<td>+17/14</td>
<td>+17/14</td>
<td>3,767</td>
<td>381/3/15</td>
<td>280</td>
<td>+4.91%</td>
<td>43</td>
<td>9,213</td>
</tr>
<tr>
<td>AOL</td>
<td>162/13/1/63</td>
<td>163</td>
<td>163</td>
<td>+B</td>
<td>+B</td>
<td>10,080</td>
<td>164/15/2</td>
<td>158/12</td>
<td>+5.16%</td>
<td>280</td>
<td>5,705</td>
</tr>
<tr>
<td>CMGI</td>
<td>57/9/16</td>
<td>57/1/2</td>
<td>57/1/6</td>
<td>+5/7/8</td>
<td>+5/7/8</td>
<td>1,333</td>
<td>93/12/3</td>
<td>93</td>
<td>+6.41%</td>
<td>705</td>
<td>11,700</td>
</tr>
<tr>
<td>SPLN</td>
<td>33/3/163</td>
<td>33/15/2</td>
<td>33/13/16</td>
<td>+7/16</td>
<td>+7/16</td>
<td>300/200</td>
<td>34/3/4</td>
<td>33/5/8</td>
<td>+1.31%</td>
<td>88</td>
<td>6,763</td>
</tr>
<tr>
<td>BEAS</td>
<td>13/1/2</td>
<td>13/5/8</td>
<td>13/5/6</td>
<td>-7/16</td>
<td>-7/16</td>
<td>359/200</td>
<td>13/1/2</td>
<td>11/6</td>
<td>-3.11%</td>
<td>-44</td>
<td>1,363</td>
</tr>
<tr>
<td>GNET</td>
<td>102</td>
<td>103/3/161</td>
<td>101/5/16</td>
<td>+5/18</td>
<td>+5/18</td>
<td>307/600</td>
<td>108/97</td>
<td>97</td>
<td>+6.43%</td>
<td>513</td>
<td>10,131</td>
</tr>
<tr>
<td>RNWK</td>
<td>67</td>
<td>67/1/4</td>
<td>67</td>
<td>+2/3/4</td>
<td>+2/3/4</td>
<td>1,233</td>
<td>69</td>
<td>64/15/6</td>
<td>+4.26%</td>
<td>275</td>
<td>6,700</td>
</tr>
<tr>
<td>MSFT</td>
<td>173/1/8</td>
<td>173/1/4</td>
<td>173/5/6</td>
<td>+1/34</td>
<td>+1/34</td>
<td>13,284</td>
<td>174/7/16</td>
<td>170</td>
<td>-1.02%</td>
<td>175</td>
<td>17,331</td>
</tr>
<tr>
<td>INTC</td>
<td>133/34/4</td>
<td>133/13/11/3</td>
<td>133/13/10</td>
<td>-3/16</td>
<td>-3/16</td>
<td>6,094</td>
<td>137/12/13</td>
<td>133/30</td>
<td>-2.20%</td>
<td>-525</td>
<td>26,763</td>
</tr>
<tr>
<td>TOTAL</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>205,302</td>
<td>80,993</td>
<td>2,293</td>
<td>+1.63%</td>
<td>2,938</td>
<td>143,280</td>
</tr>
</tbody>
</table>

Sparklines inspired “word size visualizations”

- charts or graphs tightly integrated into text or even computer code

Although Tufte is said to have invented sparklines, in actuality he invented only the name and popularized it as technique. Sparklines are a condensed way to present trends and variation, associated with a measurement such as average temperature or stock market activity, often embedded directly in the text; for example: The Dow Jones index for February 7, 2006. These are often used as elements of a small multiple with several lines used together. Tufte explains the sparkline as a kind of "word" that conveys rich information without breaking the flow of a sentence or paragraph made of other "words" both visual and conventional. To date, the earliest known implementation of sparklines was done by interaction designer Peter Zelchenko and programmer Mike Medved in early 1998.
According to Tufte (pg. 51):

- Graphical excellence is the well-designed presentation of interesting data
  - a matter of **substance**, **statistics**, and **design**

- Graphical excellence consists of complex ideas communicated with:
  - **clarity**, **precision**, and **efficiency**

- Graphical excellence is that what gives the viewer:
  - the **greatest number of ideas**
  - in the **shortest time**
  - with the **least ink**
  - in the **smallest space**

- Graphical excellence is nearly always multivariate

- Graphical excellence requires telling the truth about the data

(Nevertheless, visualizations should be visually pleasing and may very well have an artistic touch)
The Need for Visualization: Anscombe Quartet

Visualization of statistics results is important

<table>
<thead>
<tr>
<th>I</th>
<th>II</th>
<th>III</th>
<th>IV</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x$</td>
<td>$y$</td>
<td>$x$</td>
<td>$y$</td>
</tr>
<tr>
<td>10</td>
<td>8.04</td>
<td>10</td>
<td>7.46</td>
</tr>
<tr>
<td>8</td>
<td>6.95</td>
<td>8</td>
<td>6.77</td>
</tr>
<tr>
<td>13</td>
<td>7.58</td>
<td>13</td>
<td>12.74</td>
</tr>
<tr>
<td>9</td>
<td>8.81</td>
<td>9</td>
<td>7.11</td>
</tr>
<tr>
<td>11</td>
<td>8.33</td>
<td>11</td>
<td>7.81</td>
</tr>
<tr>
<td>14</td>
<td>9.96</td>
<td>14</td>
<td>8.84</td>
</tr>
<tr>
<td>6</td>
<td>7.24</td>
<td>6</td>
<td>6.08</td>
</tr>
<tr>
<td>4</td>
<td>4.26</td>
<td>4</td>
<td>5.39</td>
</tr>
<tr>
<td>12</td>
<td>10.84</td>
<td>12</td>
<td>8.15</td>
</tr>
<tr>
<td>7</td>
<td>4.82</td>
<td>7</td>
<td>6.42</td>
</tr>
<tr>
<td>5</td>
<td>5.68</td>
<td>5</td>
<td>5.73</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Property</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean of $x$ in each case</td>
<td>9 (exact)</td>
</tr>
<tr>
<td>Sample variance of $x$ in each case</td>
<td>11 (exact)</td>
</tr>
<tr>
<td>Mean of $y$ in each case</td>
<td>7.50 (to 2 decimal places)</td>
</tr>
<tr>
<td>Sample variance of $y$ in each case</td>
<td>4.122 or 4.127 (to 3 decimal places)</td>
</tr>
<tr>
<td>Correlation between $x$ and $y$ in each case</td>
<td>0.816 (to 3 decimal places)</td>
</tr>
<tr>
<td>Linear regression line in each case</td>
<td>$y = 3.00 + 0.500x$ (to 2 and 3 decimal places, respectively)</td>
</tr>
</tbody>
</table>

Same statistics
Very different data
Outliers can have a significant effect on analysis
Age-Adjusted Cancer Rates (by County)

21,000 numbers
3056 counties
7 numbers per county:
- size (4)
- location (2)
- cancer rate (1)

1950-1969
Galaxy Maps

divide sky into
1,024 x 2,222 rectangles

tone = number of galaxies per rectangle
Space Debris Map (1990)

7,000 objects > 10 cm
doubles every 5 years
Train Schedule: Paris – Lyon, 1880s

plots 6 variables: army size, 2D location, direction vector, temperature, time
Rage Fear Graph: Expressive Glyphs
Chernoff Faces: Multi-Variable Display
Chernoff Faces
Graphical Display: History

- Can be more precise and revealing than numerical display
  - example: Anscombe’s quartet (pg. 13/14)
  - example: cholera map of central London, 1854, by Dr. John Snow (pg. 24)

- Can capture a large amount of information in a very small space (billions of bits on one page)
  - example: data maps for cancer incidence (pg. 17)
  - example: galaxy maps (pg. 27)
  - example: space debris (pg. 48, Tufte “Envisioning Information”)

- Can extend to time-series display
  - example: train schedule Paris-Lyon, 1880s (pg. 31)

- Can be narrative
  - example: Napoleon’s Russia campaign, 1812, plots 6 variables on a 2D graph (pg. 41)

- Can represent each datapoint by visual information (graphic, icon, image, color, pattern)
  - examples: fear-rage graph (pg. 50), Chernoff faces (pg. 97, 142)
Tufte’s views on

- visual embellishments → “chart junk”
- abuse of physically-motivated distortions → “lie factor”
Avoid Misleading Embellishments = Chart Junk
Avoid Misleading Scaling
Manipulation of Axis Orientation

Gun deaths in Florida

Number of murders committed using firearms

2005 Florida enacted its ‘Stand Your Ground’ law

Source: Florida Department of Law Enforcement

from Panday at al. (CHI 2015)
Avoid Misleading Scaling

Commission Payments To Travel Agents
In millions of dollars

- 1976: $70
- 1977: $79
- First Half 1978: $84
- Second Half 1978: $100
- 1979: $102
- 1980: $109

Logos of airlines: Delta, Eastern, TWA, United Airlines
Avoid Misleading Use of Graphics Effects

This line, representing 18 miles per gallon in 1978, is 0.6 inches long.

This line, representing 27.5 miles per gallon in 1985, is 5.3 inches long.

real effect: \[
\frac{27.5 - 18}{18} = 53 \%
\]

graphical effect: \[
\frac{5.3" - 0.6"}{0.6"} = 783 \%
\]  
\[\Rightarrow \text{lie factor: } \frac{783}{53} = 14.8\]
Tell the Truth About the Data

REQUIRED FUEL ECONOMY STANDARDS:
NEW CARS BUILT FROM 1978 TO 1985

- 1978: 18 mpg
- 1979: 19 mpg
- 1980: 20 mpg
- 1981: 22 mpg
- 1982: 24 mpg
- 1983: 26 mpg
- 1984: 27 mpg
- 1985: 27.5 mpg

19.1 mpg, expected average for all cars on road, 1985

13.7 mpg, average for all cars on road, 1978
If You Must Embellish…

REQUIRED FUEL ECONOMY STANDARDS:
NEW CARS BUILT FROM 1978 TO 1985

19.1 mpg, expected average for all cars on road, 1985

13.7 mpg, average for all cars on road, 1978

Avoid Suggestive Distortions

IN THE BARREL...
Price per bbl. of light crude, leaving Saudi Arabia on Jan. 1

April 1
$14.55

1973  $0.95
1974  $2.41
1975  $4.36
1976  $10.46
1977  $11.51
1978  $12.09
1979  $13.34
1979  $14.55

Show the Data in Their Proper Context

Connecticut Traffic Deaths, Before (1955) and After (1956) Strict Enforcement by the Police Against Cars Exceeding Speed Limit

A few more data points add immensely to the account:

Traffic Deaths per 100,000 Persons in Connecticut, Massachusetts, Rhode Island, and New York, 1911-1959
Avoid Display of Out-of-Context Data

SOLAR RADIATION AND STOCK PRICES

Graphical Excellence

- Is cosmetic decoration really needed to make data more interesting (may only distract):
  - example: diamond graph (adds a useless 3rd dimension)

- Misleading graphical representation
  - example: missing baseline in Day Mines, Inc. annual report (pg. 54)
  - example: non-uniform data spans in Commision Payments graph (pg. 54)
  - example: non-uniform scaling of icons in Pittsburgh Civic Commission report (pg. 55)

- The Lie Factor = \( \frac{\text{size of effect shown in graphic}}{\text{size of effect in data}} \) (should be within [0.95, 1.05])
  - example: graph on fuel economy standards for autos (lie factor = 14.8) (pg. 57)

- Visualizing data bearing some dimension by means of objects of higher dimensions:
  - example: the *growing barrel* (lie factor: 9.4 (2D), 59.4 (3D)) (pg. 62)
  - example: the *growing oil pump* (lie factor: 9.5) (pg. 62)
  - example: the *shrinking dollar bill* (lie factor: ~6) (pg. 70)
  - example: the *incredibly shrinking family doctor* (pg. 69)

→ the number of information carrying dimensions should not exceed the data dimensions
• Quoting data out of context and/or too sparse (recall: graphics allows high data density)
  - example: Connecticut traffic deaths (pg. 74/75)

Principles that ensure graphical integrity:

• The representation of numbers should be directly proportional to the numerical quantities represented (see the growing barrels)
• Clear and detailed labeling should be used to defeat graphical distortion and ambiguity
• Show data variations and not design variations (see the fuel economy graph)
• In time-series displays of money, show deflated and standardized units
• The number of information carrying dimensions should not exceed the data dimensions (see the growing barrels, the shrinking doctor)
• Graphics must not quote data out of context (see the Connecticut traffic deaths)
• Convincing graphics must demonstrate cause and effect (see Challenger disaster)
Do these bare graphs engage a human audience?
  • are they memorable?

A recent (research) trend
  • will embellishment help memorability, engagement?
  • do we need what Tufte calls “chart junk”
Memorability Experiment by Borkin et al.

Experiment set up as a game on Amazon Mechanical Turk

- workers were presented with a sequence of images (about 120)
- presented for 1 second, with a 1.4 second gap between consecutive images
- workers had to press a key if they saw an image for the second time in the sequence (spacing 1-7 images with “filler” images in between)
Memorability Experiment by Borkin et al.

most memorable

most memorable after removing human recognizable cartoons

least memorable

Borkin et al. IEEE TVCG 2014
What Do People Remember?

Experiment Design

Labeled Visualization Database

393 visualizations

Visualizations are taken from [8], and the label taxonomy described in Table 1 is applied.

Encoding

100 “target” visualizations

10 seconds / image

Output: Eye-tracking fixation locations and durations.

Recognition

Same 100 targets + 100 “fillers”

2 seconds / image

Output: Eye-tracking fixation locations and durations, and whether visualization is recognized.

Recall

Correctly recognized blurred targets

20 min - as many images as participant can complete

Output: Text descriptions of what participant recalls about the visualization.

Borkin et al. IEEE TVCG 2016
Fig. 7. Examples of the most and least recognizable visualizations from [8]. **TOP:** Eye-tracking fixation heat maps (i.e., average of all participants' fixation locations) from the *encoding* phase of the experiment in which each visualization was presented for 10 seconds. The fixation patterns demonstrate visual exploration of the visualization. **BOTTOM:** Eye-tracking fixation heat maps from the *recognition* phase of the experiment in which each visualization was presented for 2 seconds or until response. The most recognizable visualizations all have a single focus in the center indicating quick recognition of the visualization, whereas the least recognizable visualizations have fixation patterns similar to the encoding fixations indicative of visual exploration (e.g., title, text, etc.) for recognition.

Borkin et al. IEEE TVCG 2016
Takeaways:

- 393 visualizations and eye movements of 33 participants and 1,000s of participant-generated text descriptions of the visualizations
- titles and supporting text should convey the message of a visualization
- if used appropriately, pictograms do not interfere with understanding and can improve recognition
- redundancy helps effectively communicate the message
- visualizations that are memorable “at-a-glance” are also capable of effectively conveying the message of the visualization
  → thus, a memorable visualization is often also an effective one

Borkin et al. IEEE TVCG 2016
Important for Memorability

Important are:

• attributes like color
• inclusion of a human recognizable object

However, link to human engagement not explicitly established

• “just” memorability

Our own studies show that embellishments can get humans interested in studying an image

• but prefer conventional charts for problem solving
### Visualizations Sources and Origins

#### Percent of Visualization Source by Visualization Type

<table>
<thead>
<tr>
<th>Visualization Type</th>
<th>Bars</th>
<th>Table</th>
<th>Diagrams</th>
<th>Lines</th>
<th>Maps</th>
<th>Points</th>
<th>Area</th>
<th>Circles</th>
<th>Trees and Networks</th>
<th>Distribution</th>
<th>Grid and Matrix</th>
<th>Text</th>
</tr>
</thead>
<tbody>
<tr>
<td>% of all Scientific Publications</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0%</td>
<td>10%</td>
<td>20%</td>
<td>30%</td>
<td>40%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0%</td>
<td>10%</td>
<td>20%</td>
<td>30%</td>
<td>40%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0%</td>
<td>10%</td>
<td>20%</td>
<td>30%</td>
<td>40%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0%</td>
<td>10%</td>
<td>20%</td>
<td>30%</td>
<td>40%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Publication Information

Borkin et al. IEEE TVCG 2014
Infographic

Graphic visual representations of information, data or knowledge intended to present information quickly and clearly

Evolved in recent years to be for mass communication
- designed with fewer assumptions about the readers knowledge base than other types of visualizations
- but can be misleading and express the opinion of the author
Using Icons as Bar Graphs

Wang et al. CHI 18
Data-Driven Design Guides

Kim et al. TVCG 17
MemViz: A Tool for Creating Memorable Visualizations

Darius Coelho, Sungsoo Ha, Shenghui Cheng, Salman Mahmood, Jisung Kim, and Klaus Mueller

Visual Analytics and Imaging Lab, Computer Science Department, Stony Brook University and SUNY Korea